Sequential observer selection for source localization

Sabina Zejnilović*†, João Gomes† and Bruno Sinopoli*
*Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA
†Institute for Systems and Robotics, Instituto Superior Técnico, Universidade de Lisboa, Portugal

***Abstract*—Identifying the source of network diffusion is an important task in applications such as epidemics management and understanding the trend propagation over social networks. As observing each node carries a cost, we study the problem of sequential selection of observed nodes from two aspects: which nodes to observe such that the source is localized with the lowest cost, and for a pre-specified number of time-steps, which nodes to observe such that the resulting number of possible source candidates is the lowest. We show that both problems can be framed, under a simple propagation scenario, as dynamic programing with imperfect state knowledge. The proposed approach is optimal, but computationally intensive, hence we propose two simple greedy strategies. Using adaptive submodularity, we provide performance guarantees for one greedy algorithm. We evaluate the proposed approaches through simulation.**

***Index Terms*—network theory, source localization, dynamic programming, adaptive submodularity**

I. Introduction

Propagation of different phenomena over networks can be modeled as network diffusion. Examples vary from spreading of viruses in human populations to information diffusion in social networks. The tasks of understanding the origin of disease, curbing infections, or determining influential individuals, all rely on source localization [1]–[6].

Source localization can be performed based on the times when nodes became "infected". However, due to network size, limited resources and privacy issues, the infection times cannot be observed for all the nodes [2], [3], [6]. The choice of the observed nodes, denoted as the observers, strongly influences the performance of the source estimator and thus selecting the most informative subset becomes an important task. The performance of high-degree nodes is compared to randomly selected nodes through simulation in [2]. Selection strategies, based on different centrality measures, are experimentally evaluated in [6]. In [7], the problem of finding the smallest subset of observers to achieve correct source localization, under a simple deterministic propagation model, is formulated as the problem of finding the smallest resolving set. In a random setting, the Chernoff distance is used as a metric in [8] to select observers that yield the lowest source localization error, when incubation times are modeled as exponential noise. In all the previously discussed strategies, all observers were selected at the same time. In this paper, we analyze the selection strategy when nodes are chosen dynamically, as the current observer is selected based on the infection times of the previous observers. This might be useful, for example, when a person who initiated a certain trend over a social network should be identified. Then the choice of which blog or site should be examined to help track this person is made after reading the previous site, as the newly acquired information is used to narrow down the search.

In order to theoretically analyze the problem to gain insight for more realistic scenarios, we apply a simple propagation model where the infection times correspond to the nodes' graph distances to the unknown source. We examine the dynamic selection problem from two perspectives. First, we wish to find a selection strategy such that the source can be unambiguously localized with the smallest cost. The second problem we analyze is when the the number of nodes that can be observed is predefined and we look at the strategy that would result with the smallest number of source candidates. We show that both these problems can be optimally solved using dynamic programming with imperfect state knowledge. However, since most networks of interest are very large, the computation cost of the optimal approach is prohibitive. Hence, we propose two efficient approximation strategies and illustrate their performance using small examples. Using adaptive submodularity, we can show that one proposed approximation algorithm has near-optimal performance guarantees [9].

II. Model setup and problem statement

We assume a widely accepted Susceptible-Infected propagation model, where once a node is infected (informed), it remains as such [1], [2], [4]. Initially, there is only a single infected/informed node in the network, the source node. At a known time, assumed 0, the source node initiates the network diffusion, modeling the scenarios when some known external event triggers propagation. We will assume the network to be a connected and undirected graph, as infections and rumors spread through contact and ties which are typically bidirectional. We adopt a simple model of diffusion where, once a node is infected at $t-1$, in the next time instant t, where t is a discrete time index, it will infect all of its neighbors, with probability 1. Then, the time of infection of a node corresponds to its graph distance to the unknown source node. Monitoring nodes is costly in terms of resources, time and effort, hence, only a subset of nodes can be observed and the source is identified using the infection times of observed nodes. A network of n nodes is represented using a graph

This research was partially supported by Fundação para a Ciência e a Tecnologia (project FCT [UID/EEA/50009/2013] and a PhD grant from the Carnegie Mellon-Portugal program) and EU FP7 project MORPH (grant agreement no. 288704).

$G = \{V, E\}$, where $V = \{1, \ldots, n\}$ is the set of nodes and E is the set of edges. The distance between two nodes i and j, $d(i, j)$, in a connected graph is the number of edges in the shortest path between them. If $O \subseteq V$ denotes the set of nodes $\{o_1, \ldots, o_p\}$, then $\boldsymbol{d}(i, O)$ is the p-vector of distances $[d(i, o_1), \ldots, d(i, o_p)]^T$. In [7], it was shown that any source can be unambiguously identified only if the set of observers forms a resolving set. In a resolving set of nodes O, we have $\boldsymbol{d}(i, O) \neq \boldsymbol{d}(j, O)$ for any two different nodes $i, j \in V$ [10]. Finding a resolving set of minimum cardinality is NP hard for a general graph, and the value of minimum cardinality is known as the metric dimension.

In order to observe a node i, a cost $c(i)$ is incurred. We wish to sequentially select nodes to localize the source at the lowest cost. This cost can reflect the time-steps needed for source localization, in which case the cost is equal for all the nodes, or it can differ between nodes, as more effort is required for accessing certain nodes. A source s can be unambiguously identified by an observer set O if and only if $\boldsymbol{d}(s, O) \neq \boldsymbol{d}(i, O)$, for all nodes $i \in V, i \neq s$. However, as the identity of node s is unknown, the goal is to find a strategy π that minimizes the expected cost, for all possible sources $s \in V$. We formulate the problem of determining the strategy π that maps the infection times of previous observers to the choice of the next observer as

$$\min_{\pi} \mathbb{E}_s \left[c\left(O\left(\pi \right) \right) \right]$$
$$\text{subject to } \boldsymbol{d}(O\left(\pi \right), s) \neq \boldsymbol{d}(O\left(\pi \right), i), \forall s \in V, s \neq i, \quad (1)$$

where $O\left(\pi \right)$ is the set of observers selected according to strategy π and $c(O\left(\pi \right))$ is the sum of costs of all nodes in the set. Let $\boldsymbol{t}_{inf}$ be a vector of observed infection times. We denote with $S(O) = \{s_1, ..., s_l\}$ a set of source candidates after a set O has been observed, i.e., $S(O) = \{s : \boldsymbol{d}(O, s) = \boldsymbol{t}_{inf}\}$. It is important to note that the members of set O are selected one at a time, and the selection stops when $S(O) = \{s\}$, i.e., there is only a single source candidate. The order by which the nodes are selected influences the total cost $c(O)$, as for different sequences of observers the stopping criterion might be met at different times, thereby incurring different total cost.

On the other hand, sometimes actions need to be taken after a pre-specified number of time steps T, equal to the number of sequentially chosen observers, even if there is some ambiguity in the source identity. Then the goal is to find a strategy π for observer selection such that the expected ambiguity is the lowest, i.e., the expected number of source candidates is the smallest. The problem can be stated as follows

$$\min_{\pi} \mathbb{E}_s |S\left(O\left(\pi \right) \right)|$$
$$\text{subject to } |O\left(\pi \right)| \leq T, \quad (2)$$

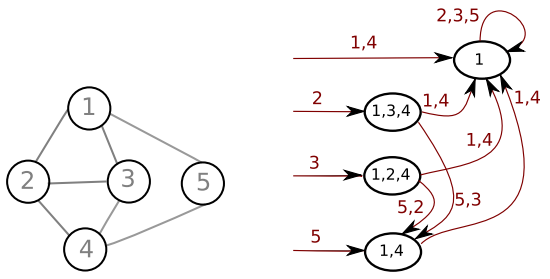
where again the expectation refers to all the possible sources.

III. Dynamic selection strategies

Stochastic dynamic programming is an optimization methodology for problems where information becomes available sequentially, and after new information becomes available, a certain action is selected [11]. The state of the system corresponds to the identity of the source. Each node has a certain probability of being the source, encoded as the prior distribution over the nodes. When prior is not available, these are set to $1/N$. The identity of the source cannot be directly observed, and instead the distances to it are known. We will model the available information about the source as the identity of the nodes that have the same distance vector to the selected observers. We denote as $S_k = \{s_1, \ldots, s_l\}$ the source candidates after k nodes have been observed. Figure 1b shows an analysis of resulting source candidates for all possible observer sequences, when the source is node 1 and the network is as shown on 1a. At the beginning of each time-step k, the information vector based on which the subsequent action is taken is $I_k = (o_1, \ldots, o_{k-1}, S_{k-1})$. The selection of observer o_k represents the action that is taken at time step k. Depending on the different sequences of selected observers, the goal might be reached at different time-steps. If the source is identified at the end of time-step k, i.e., $|S_k| = 1$, we set the cost g of each subsequent action to 0, i.e., $g_l(o_l = i) = 0$, for $l > k$ and $\forall i \in V$. Until the source is identified, the cost of selecting each node i is its cost $c(i)$, i.e., $g_l(o_l = i) = c(i)$, for $l \leq k$, if the source is identified in the $k-$th step. We set the number of time steps equal to the *resolving number* of the graph, r. The resolving number is the minimum number p such that every p-subset of V is a resolving set of G [12]. Unlike metric dimension, it can be determined in polynomial time. After r time-steps, every observer sequence results in a single source candidate, and there is no need to analyze longer sequences. The terminal cost for all sequences is set to 0. The tail cost $J_r(I_r)$ for the last step is calculated as [11]

$$J_r(I_r) = \min_{o_r \in V} \left[\mathbb{E}_s \left\{ g_r(o_r) | I_r, o_r \right\} \right], \quad (3)$$

where the expectation is taken over all the possible sources $s \in V$. For each possible information vector I_r, which contains all



(a) An example network

(b) Observation diagram when node 1 is the source

Fig. 1: Analysis of source candidates (b) for a network shown in (a). The source candidates are shown given that the true source is node 1. Red arrows and numbers on (b) represent the nodes selected for observation, while the ovals show the resulting source candidates, with some self-loops omitted. For example, if node 4 is selected as the first observer, the source is exactly determined, as $S(\{4\}) = \{1\}$, while selecting the sequence of observers $5, 2, 3$ results in source candidates $1, 4$, since $S(\{5, 2, 3\}) = \{1, 4\}$. The metric dimension of the network is 2, while the resolving number is 4.

preceding sequences of $r-1$ observers and the resulting source candidates of each sequence, the optimal cost (3) is evaluated, and the observer that achieves this cost is selected as the last observer. For the preceding observers $k = 1, \ldots, r-1$, the tail cost is given as the solution of the optimization problem

$$J_k(I_k) = \min_{o_k \in V} \left[\mathbb{E}_s \left\{ g_k(o_k) + J_{k+1}(I_k, o_k, S_k) | I_k, o_k \right\} \right]. \quad (4)$$

Now the tail cost at time k not only includes the cost of selecting each node as the observer, but also the remaining cost-to-go J_{k+1}. Again, for each possible information vector, a node that minimizes the cost at step k is chosen as the optimal observer. The information vector I_1 is empty, as there is no observation available for the selection of the first observer. The obtained optimal sequences are of length r, but only the k first nodes are of interest, where k is the first step for which $|S_k| = 1$. The calculations (3) and (4) can be done off-line, before the observer selection starts.

The value of the optimal cost J_1 represents the total optimal cost, an expected amount of resources that are spent to identify the source without any uncertainty. If the cost of all the nodes equals 1, the optimal cost is upper bounded by the metric dimension. The reasoning is as follows: one selection strategy could be observing the nodes that form a resolving set. With this strategy, it is guaranteed that regardless of the source, there would be no ambiguity after selection of such a set. Now, equations (3) and (4) give an optimal solution which cannot be larger than the solution produced by any other strategy.

In order to apply dynamic programming to optimally solve optimization problem (2), we slightly modify the previously described setup. Now, the horizon is set to T time-steps, and different sequences of the same length have different terminal cost. The cost is associated with the cardinality of the set of candidates; the larger the set is, the larger the uncertainty, and therefore, the T-th observer is selected as

$$J_T(I_T) = \min_{o_T \in V} \left[\mathbb{E}_s \left\{ |S_T| | I_T, o_T \right\} \right]. \quad (5)$$

The remaining steps, $k = 1, \ldots, T-1$ do not add any additional cost, as the objective is the smallest uncertainty after all the steps have been taken. Hence the preceding tail-costs only average over possible sources as follows

$$J_k(I_k) = \min_{o_k \in V} \left[\mathbb{E}_s \left\{ J_{k+1}(I_k, o_k, S_k) | I_k, o_k \right\} \right]. \quad (6)$$

Again, these calculations can be completed before the selection process starts. After the first node is selected, based on its infection time, the subsequent observer is chosen as the node that minimizes the cost J_2, for the given information vector. From (6) the obtained optimal cost, J_1 represents an expected number of node suspects after observing T observers. The results of applying (4) and (6) to a network depicted in Figure 1a is shown in Table I.

Even though dynamic programming leads to optimal solutions, it is not a feasible strategy for larger networks. Problems (1) and (2) are of combinatorial nature and there is an exponential growth of computational and storage requirements as the network size increases. Hence, we need to resort to

TABLE I: Expected cost for being selected as o_1 for each node of the network shown in Figure 1a evaluated with (4) *and* (6)*, assuming all nodes have cost equal to* 1 *and* $T = 1$*. Optimal* J_1 *is the smallest such cost. For problem* (1)*, the expected cost is calculated as* $\mathbb{E}_s \left\{ 1 + J_2(o_1, S_1) | o_1 \right\}$*, with* $J_1 = 8/5$ *and any one of the nodes* $1, 2, 3$ *or* 4 *can be selected as the optimal first observer. As for problem* (2) *when only one can be selected for observation, the expected cost is* $\mathbb{E}_s \left\{ |S_1| | o_1 \right\}$*, the optimal node is node* 5*, and* $J_1 = 9/5$*, i.e., on average there will be* $9/5$ *source candidates after observing the infection time of node* 5*.*

Expected cost of o_1 for \ Node	1	2	3	4	5
Problem (1) from (4)	$8/5$	$8/5$	$8/5$	$8/5$	$9/5$
Problem (2) from (6)	$11/5$	$11/5$	$11/5$	$11/5$	$9/5$

sub-optimal, yet more efficient, selection strategies. However, we can reformulate (1) and (2) with uniform prior as an adaptive stochastic optimization problem to take advantage of guarantees available for its greedy approximate algorithms which obtain near-optimal solutions [9].

Problem (1) can be cast as an *Adaptive Stochastic Minimum Cost Cover* problem as follows

$$\min_{\pi} \ \mathbb{E}_s \left[c\left(O\left(\pi \right) \right) \right]$$
$$\text{subject to } \ N - S\left(O\left(\pi \right) \right) \geq N - 1, \forall s \in V, s \neq i, \quad (7)$$

while (2) can be stated as *Adaptive Stochastic Maximization*

$$\max_{\pi} \ \mathbb{E}_s \left[N - \left| S\left(O\left(\pi \right) \right) \right| \right]$$
$$\text{subject to } \ \left| O\left(\pi \right) \right| \leq T. \quad (8)$$

Adaptive submodular functions are a generalization of submodular set functions to adaptive policies [9]. Whereas submodularity means that the benefit of an item when added to a set will not increase compared to adding the same item to its subset, adaptive submodularity reflects that selecting an item later in a sequence will not increase its *expected* marginal benefit, where expectation is computed with respect to the posterior probability given the current observations.

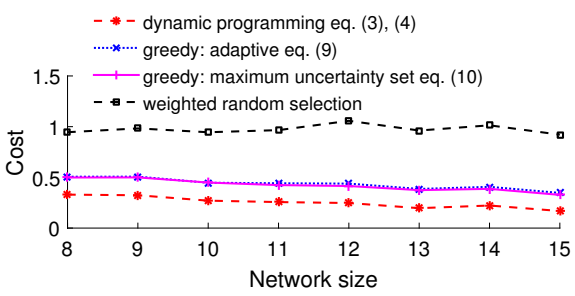
Claim: For a uniform source prior, function $f(s, O) = N - |S(O)|$ is adaptive monotone and submodular.

We omit the proof due to space limitations.

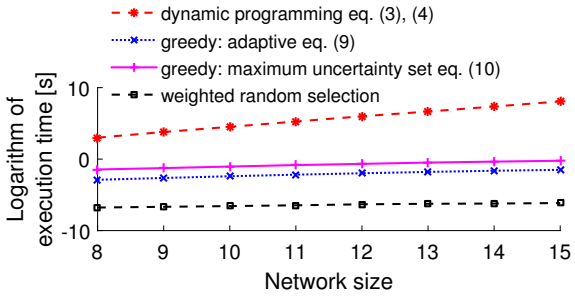
Now, leveraging on the adaptive submodularity property of function f, we apply a greedy approximation algorithm. Let $O = \{o_1, \ldots, o_{k-1}\}$ be a set of nodes selected in the first $k-1$ steps and let $I_k = (O_{k-1}, S(O_{k-1}))$. Then, at the beginning of time-step k, we choose the observer o_k as

$$o_k = \arg\max_{o \in V} \frac{1}{c(o)} \mathop{\mathbb{E}}_{s \in S(O_{k-1})} \left[|S(O_{k-1})| - |S(O_{k-1} \cup o)| I_k \right]. \quad (9)$$

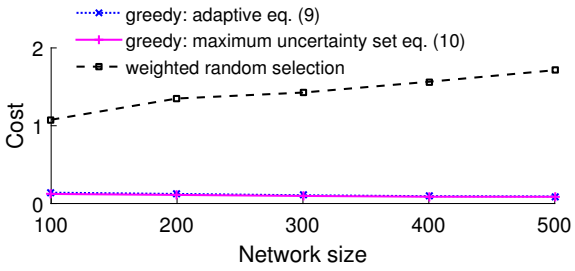
Hence, at each step an observer node is selected that minimizes the weighted expected number of source candidates, where expectation is taken considering only the current source candidates. The node weight is inversely proportional to its cost. In order to solve (7), we repeat step (9) until $|S(O_k)| = 1$. For problem (8), we repeat the selection step



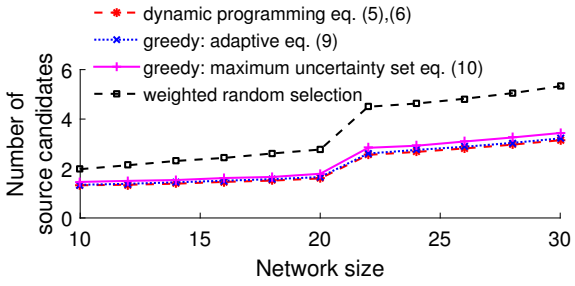
(a) Cost incurred for solving (1)



(b) Time required to solve (1)



(c) Cost incurred for solving (1) by approximate algorithms



(d) Number of source candidates after solving (2) for $T = 2$.

Fig. 2: The performance of dynamic programming and greedy approaches for solving (1) *and* (2).

a predefined number of times, setting $c(o) = 1$ for all the nodes. Let $c(\pi^{opt})$ denote the optimal cost of (7), and $c(\pi^g)$ represent the expected cost achieved by strategy (9). Then $c(\pi^g) \leq c(\pi^{opt})\,(\log(N(N-1))+1)$ is guaranteed to hold [9]. Also, let $f(\pi_k^{opt})$ denote the optimal expected value of (8) after k steps, and $f(\pi_k^g)$ represent the expected value achieved by strategy (9). Then the results in [9] guarantee that $f(\pi_k^g) \geq (1-e^{-1}) f(\pi_k^{opt})$.

We additionally propose a third strategy for observer selection that at each step minimizes the maximum number of possible source candidates. We denote as $T_S^c(o)$ the nodes of set S which are at distance c to node o, where $c = 1, \ldots, l$, and l is the maximum such distance, i.e. $T_S^c(o) = \{t : t \in S,\, d(o,t) = c\}$. Then, at the beginning of step k, with S_{k-1} defined as previously, observer o_k is chosen as

$$o_k = \arg\min_{o \in V} c(o) \max_c |T^c_{S_{k-1}}(o)|. \tag{10}$$

Unlike in the adaptive algorithm (9) where an observer is selected that minimizes the weighted *expected* number of source candidates, here at each step an observer is selected based on the worst case scenario, by minimizing the weighted *highest* possible number of source candidates.

To illustrate the merits of the proposed greedy approaches we will be comparing their performance to a weighted random selection that selects more costly nodes with less probability. Specifically, this random selection that is used as a benchmark at each step randomly selects a node i with normalized probability $p(i)$ inversely proportional to its cost, i.e., $p(i) = \frac{1}{c(i)} / \sum_j \frac{1}{c(j)}$.

IV. Simulation results

We illustrate the performances of the proposed approaches for randomly generated small world networks. For each network size, 100 realizations were considered, a uniform source prior was used, and node costs were chosen randomly in the range $(0, 1)$. Figure 2a shows that for small networks, the greedy approaches yield similar average cost for (1), significantly lower than the cost incurred with random selection, and the gap between greedy performance and the optimal solutions does not change much as the network size increases. Figure 2b shows the corresponding average execution time, which is, as expected, much higher for optimal solution. For larger networks, we only show the cost attained by the greedy approaches. Figure 2c shows that the greedy approaches again have very similar performance, much lower that the random selection and the gap between greedy and random selection increases as the number of nodes increases. When the number of observed nodes is set to 2, the greedy adaptive algorithm (9) yields a slightly lower number of source candidates than (10), with both not much higher than the optimal strategy for (2), unlike the random selection, as shown in Figure 2d.

V. Conclusions

We analyzed the problem of sequential observer selection for source localization from two perspectives: minimizing the observation cost for unambiguous source localization and minimizing the number of source candidates after observing a prespecified number of nodes. We solve the problems optimally with dynamic programming, which is not efficient for large networks. We additionally propose two simple greedy approaches with similar performances, both comparable to optimal and significantly better than a weighted random selection. For one greedy approach we can show it has performance guarantees under the framework of adaptive submodularity.

REFERENCES

[1] D. Shah and T. Zaman, "Rumors in a network: who's the culprit?" *IEEE Transacations on Information Theory*, vol. 57, no. 8, pp. 5163–5181, 2011.

[2] P. Pinto, P. Thiran, and M. Vetterli, "Locating the source of diffusion in large-scale networks," *Physical Review Letters*, August 2012.

[3] W. Luo and W. P. Tay, "Estimating infection sources in a network with incomplete observations," *IEEE GlobalSIP*, pp. 301 – 304, 2013.

[4] W. Luo, W. P. Tay, and M. Leng, "Identifying infection sources and regions in large networks," *IEEE Transactions on Signal Processing*, vol. 61, no. 11, pp. 2850–2864, June 2013.

[5] C. H. Comin and L. da Fontoura Costa, "Identifying the starting point of a spreading process in complex networks," *Physical Review*, vol. 84, 2011.

[6] E. Seo, P. Mohapatra, and T. F. Abdelzaher, "Identifying rumors and their sources in social networks," *SPIE Defense, Security, and Sensing*, April 2012.

[7] S. Zejnilovic, J. Gomes, and B. Sinopoli, "Network observability and localization of the source of diffusion based on a subset of nodes," *Allerton*, pp. 847– 852, 2013.

[8] S. Zejnilovic, J.Xavier, J. Gomes, and B. Sinopoli, "Selecting observers for source localization via error exponents," *ISIT*, 2015.

[9] D. Golovin and A. Krause, "Adaptive submodularity: Theory and applications in active learning and stochastic optimization," *Journal of Artificial Intelligence Research*, vol. 42, pp. 427–486, 2011.

[10] S. Khuller, B. Raghavachari, and A. Rosenfeld, "Landmarks in graphs," *Discrete Applied Mathematics*, vol. 70, no. 3, pp. 217–229, 1996.

[11] D. Bertsekas, *Dynamic programming and Optimal Control*. Athena Scientific, 1995, vol. I.

[12] D, Garijo, A. Gonzalez, and A. Marquez, "The resolving number of a graph," *arXiv:1309.0252v1*, 2013.